Research Analyse

Transfertool

Naam: Shibin Pan

Klas: Seito21A

Stagebedrijf: Prodist ERP Software

Stagebegeleider: William Cornelissen

Datum: 20/02/2024

Versie: 0.1

Inleiding

In dit document wordt een grondige research analyse gedaan voor het ontwikkelen van een geautomatiseerde tool TransferTool, die gericht is op het automatisch verwerken van PDF-bestanden en het genereren van XML-bestand, om vervolgens te kunnen importeren in ProdistERP-applicatie. Het doel van deze analyse is om inzicht te krijgen in de bestaande oplossingen, technologieën en beste praktijken die relevant zijn voor het ontwikkelen en implementeren van de tool.

In deze analyse zit verschillende aspecten, onder andere het onderzoek naar bestaande softwareoplossingen en bibliotheken, PDF-verwerkingstechnologieën, tekstherkenningsalgoritmen, machine learning en AI, gegevensbeveiliging en privacy etc.

Inhoudsopgave

[Inleiding 1](#_Toc159413925)

[Bestaande oplossingen en technologieën 3](#_Toc159413926)

[PDF-verwerkingstechnologieën 4](#_Toc159413927)

[Tekstherkenningsalgoritmen 5](#_Toc159413928)

[Machine learning en AI 6](#_Toc159413929)

[Gegevensbeveiliging en privacy 7](#_Toc159413930)

[Compatibiliteit en schaalbaarheid 8](#_Toc159413931)

[Gebruikerservaring en acceptatie 9](#_Toc159413932)

[Kostenanalyse 10](#_Toc159413933)

# Bestaande oplossingen en technologieën

Op dit moment zijn er meerdere verschillende softwareoplossingen op de markt voor het verwerken en converteren van PDF-bestanden, zoals Adobe Acrobat, Finereader en Movavi etc. Ze bieden veel functionaliteiten, onder andere het uithalen van tekst, afbeeldingen. Deze oplossingen zijn bekend om de uitgebreide mogelijkheden, maar daarbij zijn er ook hoge licentiekosten.

Voor ontwikkelomgeving zijn er ook verschillende oplossingen voor, zoals iTextsharp, PDFBox en PDF.js etc. Deze zijn bibliotheken die handig zijn voor het verwerken met PDF-bestanden in C# en andere programmeertalen. Ze bieden meerdere functionaliteiten zoals het uithalen van tekst, het bewerken van PDF’s, en genereren van nieuwe PDF-bestanden.

# PDF-verwerkingstechnologieën

iTextSharp is een van de meeste populaire bibliotheek voor het verwerken van PDF-bestanden in C#. Hij biedt uitgebreide mogelijkheden zoals, het extraheren van tekst, afbeeldingen, maar ook het maken van nieuwe PDF-bestanden, bewerken etc. Hij biedt ook ondersteuning voor andere functies zoals digitale handtekening en paginabeheer etc. Ik heb hiervoor een stukje kunnen proberen. Hij haalt letterlijk alle tekst uit een PDF-bestand, maar zonder volgorde. Het kan gebeuren dat alles door elkaar heen staat

IronPDF is ook een bekende bibliotheek. Beide heeft dezelfde functionaliteiten, onder andere het aanmaken, bewerken en manipuleren van PDF-bestande in C#. Maar IronPDF biedt een moderne API voor het werken met PDF-documenten. Verder is IronPDF een commerciële bibliotheek die wordt geleverd met een gratis proefversie en een betaalde versie.

PDFBox is een Java-bibliotheek die vergelijkbare functionaliteiten biedt als iTextSharp en IronPDF, maar het is specifiek gericht op het werken met PDF-bestanden in Java. Hij biedt verschillende mogelijkheden zoals het lezen, en genereren van PDF-bestanden, waardoor de developers zelf de controle hebben over de inhoud en opmaak van zijn PDF-document.

# Tekstherkenningsalgoritmen

OCR, staat voor Optical character recognition, of in het Nederlands: Optische tekenherkenning. Het is een manier waarbij alle tekens uit een afbeelding worden herkend door middel van patroonherkenning. Oftewel, de tekst uit een afbeelding wordt omgezet in bewerkbare tekst.

Tesseract, een open-source tekstherkenningsengine, staat bekend om tekst te herkennen in verschillende talen vanuit de afbeeldingen. Hij is zeer nauwkeurig en compatibel met andere programmeertalen en frameworks. Hij converteert dus afbeeldingdocument naar een bewerkbaar PDF of word document. Hij is wel beschikbaar voor iedereen. Maar in tegenstelling kan hij minder goed tegen handgeschreven tekst, en slechtkwaliteit gescand document.

ABBYY FineReader, aan de andere kant, is een commerciële tekstherkenningssoftware die algoritmen gebruikt voor het nauwkeurig herkennen van tekst in verschillende talen en formaten. Hij staat meer bekend om tekst uit meer complexe documenten te halen, zoals tabellen, grafieken en handgeschreven tekst. De hoge nauwkeurigheid brengt echter kosten met zich mee.

Paddle OCR is ook een open-source OCR engine. Hij kan verschillende documenttypen verwerken. Het is ontworpen om eenvoudig te gebruiken en kan worden aangepast aan specifieke usercases.

# Machine learning en AI

CNN’s, staat voor Convolutional Neutal Networks, zijn een soort deep learning model dat is zeer effectief is bij beeldherkenningstaken. OCR-oplossingen die gebruik maken van CNN’s kunnen zich met grotere nauwkeurigheid aanpassen aan verschillende lettertypen, formaten en layout. Oftewel het verbeteren van tekstherkenning in afbeeldingen.

Frameworks zoals TensorFlow en PyTorch bieden de tools en bibliotheken die nodig zijn voor het trainen van dergelijke modellen:

TensorFlow is een van de meeste populaire deep learning-frameworks. Hij ondersteunt programmeertalen zoals Python en C++ om deep learning-modellen creëren.

PyTorch is een computerframework dat brede ondersteuning biedt voor machine learning-algoritmen. In tegenstelling tot Torch. PyTorch draait op Python, wat betekent dat iedereen met een basiskennis van Python aan de slag kan met het bouwen van zijn eigen deep learning-modellen.

Zo heb je bijvoorbeeld nog meer frameworks om deep learning-modellen te bouwen, zoals DeepLearning4J voor Java. Het gebruik van machine learning vereist echter vaak uitgebreide trainingsgegevens op het gebied van modelontwikkeling.

# Gegevensbeveiliging en privacy

GDPR staat voor General Data Protection Regulation. Oftewel in het Nederlands: Algemene verordening gegevensbescherming, is een Europese verordening die de regels voor de verwerking van persoonsgegevens door particuliere bedrijven.

Bij het verwerken van gevoelige informatie zoals facturen of bedrijfsdocumenten is het belangrijk om te voldoen aan de eisen. Dat betekent dat de TransferTool mogelijk maatregelen moet implementeren zoals gegevensversleuteling, toegangscontrole en gegevensverwijdering etc, om aan deze eisen te voldoen.

Maar alle data’s zelf worden niet opgeslagen tijdens of na de overdracht. Het PDF-bestand worden uitgelezen en dan geïmporteerd naar een nieuw gemaakt XML-bestand, om dat vervolgens te kunnen importeren in ProdistERP-applicatie. Dus alles gebeurt op lokaal computer.

# Compatibiliteit en schaalbaarheid

De transferTool is alleen te werken op Windows. Een specifieke map wordt constant in de gaten gehouden, zodra er nieuwe bestanden worden gedetecteerd, dan gaat de tool van start.

De transferTool moet wel tegelijkertijd meerdere PDF-bestanden kunnen behandelen. Dus een gebruiker sleept of kopieert tegelijkertijd meerdere bestanden, en de tool moet wel voor elke bestand een check kunnen doen of het een PDF-bestand is. Zo ja, dan wordt alle gegevens voor dat bestand uitgelezen, en in XML-bestand te importeren.

# Gebruikerservaring en acceptatie

De prototype versie krijgt de gebruiker een scherm waarmee hij een/meer bestanden kan uploaden. Maar als eindresultaat kan de gebruiker gelijk via verkenner zijn bestanden uploaden, dan wordt de rest voor hem gedaan. Dus de gebruiker krijgt niet veel te zien, alleen dat nieuwe XML-bestanden dat is aangemaakt door de tool.

Maar wat wel belangrijk is dat de overdacht snel en betrouwbaar moet zijn bij het extraheren van tekst en andere relevante gegevens uit de PDF-bestanden en het importeren naar nieuwe bestand.

# Kostenanalyse

Door een gedetailleerde kostenanalyse voor het ontwikkelen uit te voeren, kunnen de kosten van het project worden bepaald.

Dit project is een stageopdracht, dus in principe is er geen ontwikkelkosten voor de ontwikkelaars. Wel is dat alles ingeleverd moet worden voor 27 mei 2024, en dan wordt het project op 4 juni 2024 gepresenteerd. Maar daarvoor zijn er zeker wel meerdere testrondes gedaan.

De opdracht is ook beperkt tot PDF-bestanden van een hotelketen. Dus de opmaak van de bestanden zien er hetzelfde uit. Maar als de tool later uitgebreider gaat worden, dus voor willekeurige PDF-bestanden, of dat hij ook voor voor andere bedrijven geldt, dan brengt er natuurlijk wel extra kosten voor het implementeren.